An Application of Statistical Matching with the Survey of Income and Education and the 1976 Health Interview Survey

Norma I. Gavin

This article outlines an alternative procedure to household surveys for obtaining individual observation-level data. The procedure, called statistical matching, integrates data on an individual observation from one source with data on a different observation identified as the "best matching" or "most similar" record from a second source. The best match is determined by objective statistical criteria. Also reported is a significant application of the procedure between the Survey of Income and Education and the 1976 National Health Interview Survey. The success of merging these two large, nationally representative data files shows statistical matching as a viable method of creating databases for health services research.

Studies have shown that not only the size but also the demographic composition, socioeconomic status, and geographic distribution of the population have a significant effect on the use of health services and the related expenditures. Therefore, the current and future characteristics of the consumer-patient population are of great concern to health planners, policy analysts, and health care providers. Population-based files of health care utilization and expenditure data on individual observations can help answer many of their questions.

The project reported in this article was conducted while the author was at Mathematica Policy Research, Inc., under contract with the Health Care Financing Administration and the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services pursuant to contract numbers HCFA-500-74-60 and HEW-100-79-0161.

Address correspondence and requests for reprints to Norma I. Gavin, M.A., Research Assistant, Center for Research of the College of Business Administration, The Pennsylvania State University, 103 BAB II, University Park, PA 16802.

Household survey files are the usual source of such data. However, surveys are time-consuming and costly and often cannot be made for every research need. Regular and periodic surveys, such as the National Health Interview Survey and those conducted by the Center for Health Administration Studies and the National Opinion Research Center at The University of Chicago, while rich with information, often lack a critical variable for a particular study.

This paper outlines an alternative procedure for obtaining the required database. The procedure, called statistical matching, is a method of integrating data on individual observations from one source with data on similar but different observations from a second source. It is used to add variables to records in a data file in which they did not exist before and to correct or adjust the distributions of variable values on a data file where there is reason to believe that the distribution on one file is superior to that on the other.

The procedure has been used by analysts to construct more comprehensive and accurate databases from existing ones for estimates of the distributions of income, taxes, wealth, energy consumption, and the costs and effects of changes in government programs [1–12]. However, its potential use in developing databases for health services research has not been fully evaluated. A pioneering statistical match between the Survey of Income and Education (SIE) and the 1976 Health Interview Survey (HIS) is described.[1] It is hoped that the presentation will provide a base for discussion on the appropriateness of statistical matching in creating databases to study certain health care issues.

A Description of the Procedure

In a statistical match, composite data records are created by selecting a record from one file that best matches a given record on a second file and appending information on that record to the record in the second file. The best match is determined by an objective statistical criterion, such as the minimization of a distance function of the differences between values of variables common to the two files or the maximization of a score where points are given for matches on the values of the common variables.

Alternative statistical techniques, such as regression analysis, can also be used to impute variables contained in one file to another file or to adjust the values of certain variables on a file. However, where a substantial number of imputations or adjustments are required and

where the joint distributions among the imputed variables are important, a statistical match may be more appropriate.

A precondition for a statistical match, then, is the existence of two or more data files, which together contain the data needed for the analysis and which have variables in common on which to base the match, and sample populations which are consistent or can be aligned.

BRINGING ABOUT THE SIE/HIS MATCH

Several years ago, the expected congressional interest in Medicaid reform and other national health insurance programs revealed the need for a database that could provide comparable estimates of eligibles, participants, and costs of the various proposed plans. A population-based data file of records on individuals representative of the U.S. population to which the various program rules could be applied in a microsimulation model was needed.

Household surveys existing at the time lacked either the required income data to determine program eligibility among the sample population or the required health care utilization and/or expenditure data to determine program costs. Furthermore, sample sizes were too small to support state-specific analyses. State estimates are important for determining the geographic distribution of gainers and losers in a new program and are particularly relevant in comparing the current Medicaid program—in which the eligibility and benefit rules vary by state—to a federalized Medicaid program with a single set of rules.

No existing database had all of these features. However, some of the variables needed for the analysis were on the SIE file, while others were on the HIS file.

The SIE, conducted by the Census Bureau in the spring of 1976, is a national survey of approximately 151,000 households designed to give reliable state estimates of the number of children 5–17 years of age living in poverty. The file contains detailed information on income and employment necessary for the determination of eligibility for medical assistance programs, as well as certain sociodemographic data, such as age, sex, education, family income, and health insurance coverage, that are comparable to information gathered in the HIS.

The HIS is conducted annually by the National Center for Health Statistics and is composed of information on a national sample of about 42,000 households. It is designed to provide national estimates of acute illnesses and injuries, disability days, and measures of health care utilization. The utilization data are needed to determine the benefit or

transfer amounts in studies of the comparative costs of medical assistance programs.

Because the two files contained several variables in common and their sample universes were the same, a statistical match was possible.

Before the statistical properties of the variables common to the two files could be examined to determine their use in a match, their definitions had to be reconciled. The alignment of the common variables on the SIE and HIS files was fairly easy. In all cases, an attempt was made to preserve the maximum possible amount of significant detail. The distributions over the variable values were nearly identical on the two files for all of the variables, except census division and state. The different geographic distributions were expected and are due to the different sampling designs of the surveys.

Reporting errors on the surveys were discovered from comparisons with administrative program data. The underreporting of income, private health insurance, and the receipt of cash welfare payments was slightly more pronounced in the HIS, while the underreporting of the receipt of Medicaid benefits was more pronounced on the SIE. These differences can also be explained from the different survey designs. The SIE is designed to obtain income and employment data; therefore, the questionnaire contains probes to elicit more accurate information relating to these items. On the other hand, the HIS is designed to elicit accurate health care utilization data, which include Medicaid benefits. The differences did not affect the match itself because they were small and because an exact match on all of the variable values was not required.

The matching of these two files allows analysts access to the best information from both surveys. The most reliable survey data on income, employment, and health care utilization are available on a single file without unduly burdening survey respondents.

THE DESIGN OF A MATCH

Among the major elements of a match are the direction, the matching unit, the variables to be used to stratify the files and those to be used in the matching algorithm (e.g., the distance function), the weights of these latter variables, and constraints imposed on the number of times a record can be used.

THE DIRECTION

The direction of the match is determined by which data file serves as the base file and which one serves as the non-base file. The unique information from the non-base file records is transferred to the base file records; the population distributions on the final merged file will be those of the base file. Therefore, the decision on the direction of the match depends on the intended use of the output file.

The SIE and HIS data files are both representative samples of the U.S. population, but only the SIE is large enough to give reliable state estimates. Because we wanted a file that gave estimates of health care utilization and program eligibility by state, the SIE was chosen as the base file and the HIS as the non-base file in our match.

THE MATCHING UNIT

Matching is performed on a single common unit or level of observation. The data files may have different units of observation, or they may have more than one common unit. When the files have no common unit, a corresponding unit must be made by combining or dividing records on one or both of the files.

Both the SIE and HIS files have multiple observation levels, with household and/or family records for each household interviewed and person records for each household member.[2] The SIE/HIS match was made at the person level; that is, person records from the HIS were chosen to match the person records in the SIE. The choice of persons for the matching unit, rather than the family, substantially increased the cost of the match. But the choice was made because, although family characteristics have a significant influence in determining the utilization of health care services by family members, the influence of personal characteristics, such as age and sex, is much greater. Furthermore, a person match allows the analyst greater flexibility in redefining filing units for modeling alternative medical assistance programs.

THE MATCH VARIABLES

Match variables are the variables common to both files on which the matching criteria are based. They are typically grouped into stratification and weighted match variables. The stratification variables define matching cells. Only non-base file records in the cell corresponding to the base file record's cell are searched to determine the best match for a record on the base file. The use of cells greatly reduces the cost of the match; in many applications, it would be exorbitantly expensive to search every non-base file record for every base file record. The

weighted match variables then are chosen from the remaining variables common to the two files and are assigned either a weight for use in a distance function or points which are awarded for matches between the two files.

Stratifying the files into cells is equivalent to giving the stratification variables extremely high weights so that the best matching records would always match on them. Ideally, the most important determinants of the variables to be appended to or adjusted on the base file serve as stratification variables.

The desired stratification variables for the SIE/HIS match were those that are the most important determinants of health care utilization and participation in medical assistance programs. However, other factors, such as the cell sizes on the HIS and the similarity of the variable distributions on the two files, important to retain the aggregate values and distributions of the utilization variables on the merged file, were also considered. Regression analysis on the major health care utilization variables in the HIS file was used to ascertain the relative importance of the potential matching variables.

The variables chosen to stratify the files were census division, age, sex, and family size. The family size categories varied across the cells. In addition, a high-low family income flag and a disability flag were used to break down further the remaining large cells of persons under 65 years of age, and all cells of persons over age 65 were stratified by the disability flag.

THE WEIGHTS

Once the stratification variables are identified, a matching algorithm or rule can be specified, and the weights or scores for the remaining common variables can be determined. In the SIE/HIS match a scoring system was used. Each match variable was assigned a certain number of points, which were accumulated in a total score for each of the HIS records in a cell if that record and the SIE record had the same variable value. The HIS record with the highest total score greater than a minimum allowable value was determined to be the best match.

The points assigned to each of the match variables was based on their relative importance in determining the information to be appended to or adjusted on the base file. To compute the points assigned to the variables in the SIE/HIS match, we ran least-squares regression equations on the HIS doctor visit variable. This variable was chosen over other utilization data on the file because it had fewer zero responses; therefore, the bias resulting from the use of the least-

squares method is smaller than it would be with the other variables. Furthermore, because much of the utilization of health care services is either induced by the physician or naturally follows visits to physicians, the use of the number of doctor visits to represent overall health care utilization is not without a theoretical base [13,14].

The independent variables of the equation included the stratification variables and a potential scored variable. A regression equation was run for each potential scored-match variable. In addition, because of widely different health care needs and motivational factors in seeking care, separate sets of equations were run for children (0–16 years old), non-aged adults (17–64 years old), and the aged (65 or more years old).

The individual variable scores then were derived from the square of the partial correlation coefficient between the potential scored variables and the doctor-visit variable in these equations, r^2. These values represent the proportional reduction of the variation in physician visits not explained by the stratification variables.

Note that this is not the only way to determine the variable weights or scores. In most previous matches, these values were determined subjectively, based on the analysts' knowledge of the relative importance of the variables [3–5,7,8,10–12]. However, other statistical criteria have also been used [1].

The resulting r^2 values for the potential match variable are shown in Table 1. They varied considerably among the three age groups. However, the most significant variable by far for all three groups was the disability flag. State of residence and the different insurance coverage variables were also among the more significant explanatory variables for the three groups.

For the most part, the r^2 values were directly translated into the individual variable scores. However, the scores for several variables that were considered more highly correlated with participation in a medical assistance or national health insurance program or dimensions of health care utilization not reflected in the doctor visit variable were adjusted to account for these relationships. Furthermore, for continuous variables, partial points were awarded if the variable values were within a tolerance range but did not match exactly.

CONSTRAINTS

Statistical matches are usually one of two types—totally constrained or totally unconstrained. In an unconstrained match, often called matching with replacement, emphasis is made on picking the best matching

Table 1: Proportional Reduction of the Variation in the Number of Doctor Visits with the Different Potential Scored-Match Variables

	Children	*Non-aged Adults*	*Aged Adults*
State	.00363	.00171	.00689
Health insurance			
No insurance flag	.00237	.00080	.00004
Type of insurance coverage	.00389	.01813	.00487
Private insurance flag	.00003	.00266	.00004
Medicaid coverage flag	.00246	.01689	.00490
Medicare coverage flag	.00002	.00088	.00004
Military insurance flag	.00004	.00034	.00000
Type of private insurance with private insurance flag	—	.00000	.00003
Medicaid beneficiary flag with Medicaid coverage flag	.00057	.00136	.00141
Disability			
Disability flag	.04066	.06564	—
Duration of disability with disability flag	—	.00579	.00042
Family characteristics			
Family income	.00107	.00228	.00000
Education of head	.00242	.00266	.00062
Family size	.00166	.00000	.00015
Number of children under six	.00011	.00267	.00000
Children under six flag	.00029	.00239	.00000
Family structure	.00068	—	—
Personal characteristics			
Age	.00451	.00012	.00015
Infant flag (age under 2)	.01209	—	—
Race	.00049	.00138	.00054
Marital status	—	.00347	.00071
Family relationship	—	.00221	.00025
Welfare recipient flag with Medicaid coverage flag	.00002	.00004	.00368

record from the non-base file for each record in the base file. Therefore, after a non-base file record is chosen to match a base file record, it is replaced in the non-base file and may be selected as the best match for any number of base file records. There is no constraint on the number of times a given non-base file record can be chosen. What may occur in these instances is that certain records are chosen many times and others are not chosen at all. Consequently, the aggregate values

and distributions of the non-base file variables on the matched file may not be the same as the values and distributions on the original non-base file.

When it is important that the aggregate totals and distributions of these variables be maintained, a constrained match, or match without replacement, is used. In this procedure, all of the records on the non-base file must be matched to base file records, and their sample weight in the merged file must equal their original sample weight. To do this requires record splitting or other, similar techniques, and some nonoptimum matches for individual records will occur to keep the whole picture unbiased.[3]

Thus, the trade-off lies between obtaining an unbiased overall match and obtaining unbiased matches for individual records and subgroups of the population. Little is known about the relative magnitudes of these biases. They are functions of the statistical properties of the two samples, the sample sizes, and the matching method [15,16].

In the SIE/HIS match, unbiased individual records were needed, because the rules of the various Medicaid reforms and national health insurance plans were to be applied at the individual record level. At the same time, it was considered important to maintain the aggregate values of the health care utilization variables found on the HIS file. Although these variables are known to contain considerable underreporting, there is no evidence to suggest that one subgroup of the population is a better or worse reporter than another [17].

The method chosen to match the SIE and HIS files was a compromise in which the records in the non-base file were partially constrained so that they would not be chosen more than a given number of times unless they were much better matches than any of the other records: penalty points were added to a record's score if it was chosen as the best match for a given number of base file records. The penalty increased the more times a record was chosen beyond the acceptable number.

RESULTS OF THE SIE/HIS MATCH

The SIE/HIS match was accomplished in two passes of the SIE file. In the first pass, almost all of the SIE records were successfully matched—we were able to find an HIS record with a score greater than the minimum allowable score for more than 99.7 percent of the SIE records. Nearly all of the unmatched records were for disabled persons, and most were for children 11–16 years of age. Geographically, almost

40 percent of the unmatched records were for persons residing in the Mountain-states region.

In a comparison of mean health care utilization figures from the matched file with those from the HIS file, we found a considerable undercount among the adult population, particularly among the aged. The reason was found to be differing distributions of disability over the three age groups in the two files. Although the files have approximately the same number of disabled persons, the SIE has more disabled children records (5.2 percent compared to 3.8 percent on the HIS), slightly fewer disabled non-aged adult records (13.4 percent compared to 14.2 percent), and considerably fewer disabled aged adult records (36.9 percent compared to 45.2 percent). The use of health care services increases significantly with age, and in each age group, the disabled utilize considerably more health care services than do the nondisabled. Therefore, in the merged file from the first-pass match, where all successful matches were forced to have identical disability codes, health care utilization was underestimated.

To correct the undercount, SIE records of nondisabled adults with low match scores were selected to be rematched with the unmatched records from the first pass. In the second-pass match, all rematched and unmatched SIE records for adults were forced to match HIS records of disabled persons. At the same time, some of the unmatched records of disabled children were forced to match HIS records of nondisabled children.

The second-pass match was successful. The frequency of matches of the scored-match variables in the final output file are shown in Table 2 by age group. Each of the variables matched on a vast majority of the records. Variables with higher numbers of points had higher percentages of matches.

But the real success of the match is judged by how well the distributions of the health care utilization measures on the HIS are reproduced on the matched file with the SIE population weights. The totals, means, standard deviations, and the incidence of non-zero values for the major utilization variables on the HIS file compared to those on the matched file are shown in Table 3.

None of the totals on the matched file differ more than 1–2 percent from the total on the HIS file. In addition, the distributions of the health care measures are almost identical on the two files as judged by their means and standard deviations, as well as by the percentages of persons with positive values.

Also compiled were detailed tables of the percentage of disabled persons, the mean number of doctor visits, and the number of hospital-

Table 2: Frequency of Matches on the Scored-Match Variables in the Final Matched File by Age Group

	Children	*Non-aged Adults*	*Aged Adults*
State	77.2	62.1	87.4
Health insurance			
No insurance flag	96.3	96.3	97.0
Type of insurance coverage	88.8	92.3	68.0
Private insurance flag	—	95.6	—
Medicaid coverage flag	97.0	99.1	92.1
Disability flag	99.5	99.0	91.6
Family characteristics			
Family income category			
Exact match	70.3	73.2	69.6
Tolerance match	12.3	11.0	14.6
Education of head	74.5	80.3	60.8
Family size			
Exact match	76.5	81.1	76.4
Tolerance match	15.4	11.6	14.8
Number of children under 6	—	90.8	—
Children under 6 flag	78.7	95.1	—
Family structure	86.7	—	—
Personal characteristics			
Age			
Exact match	39.6	25.9	12.1
Tolerance match	33.4	26.1	39.5
Infant flag (age under 2)	99.8	—	—
Race	84.5	86.8	90.8
Marital status	—	90.6	78.6
Family relationship	—	92.1	81.3
Welfare recipient flag	95.5	—	97.1

days per 100 persons on the matched file compared to the HIS file broken down by sex, more detailed age categories, health insurance coverage, census division, and state [18]. In general, these tables show that the match was successful in reproducing the distributions of the HIS health care measures on the new file. Not surprisingly, distributions broken down by stratification variables are more closely reproduced than those broken down by scored-match variables.

To the extent that the SIE gives a better distribution of sample persons over the different age groups, and given that the distributions of the health care utilization measures for these age groups are closely

Table 3: Comparison of the Major Health Care Utilization Estimates from the 1976 HIS and the Matched SIE/HIS Files by Age Group

	Children		*Non-aged Adults*		*Aged Adults*	
	HIS	*SIE/HIS*	*HIS*	*SIE/HIS*	*HIS*	*SIE/HIS*
Outpatient doctor visits						
Total number (in 000's)	179,480	181,994	505,245	504,742	122,094	120,857
Mean number	2.95	2.99	3.94	3.94	5.63	5.57
Standard deviation	7.17	6.59	9.68	9.49	13.38	13.06
Percentage of population with one or more visits	74.8	75.8	76.2	76.6	80.3	80.1
Short-term hospital stays						
Total number (in 000's)	3,824	3,833	18,318	18,570	5,290	5,171
Mean number per 100 persons	6.29	6.30	14.30	14.50	24.39	23.84
Percentage of population with one or more stays	5.4	5.5	11.7	11.7	17.8	17.6
Short-term hospital days						
Total number (in 000's)	20,025	19,850	130,967	130,454	57,710	57,104
Mean number per 100 persons	32.9	32.7	102.2	101.8	266.1	263.3
Standard deviation	285.9	305.7	546.1	532.7	292.3	295.7
Dental visits						
Percentage of population with one or more visits	50.5	51.4	51.5	52.7	29.9	30.7

reproduced on the merged file, estimates of total utilization from the matched file may be an improvement over the HIS estimates.

However, the matching procedure may have introduced other biases in the estimates and, thus, they may be less reliable. The relative sizes of these biases are unknown, but there is no evidence that the matching procedure introduced any sizable bias.

CONCLUSION

A synthetic file for health services research was created by statistically matching the 1976 HIS file with records in the SIE file. The new file consists of records for the households and persons in the SIE sample containing all of the information collected on them in that survey. In addition, each person record is linked with the health information for an individual in the HIS sample with characteristics most nearly identical to the SIE individual.[4] No significant biases in the distributions of the health care data using the SIE population weights were found. Thus, the merged file successfully combines the most reliable, nationally representative survey data on income and employment with the most reliable, nationally representative survey data on health care utilization.

This population-based file can be used to help answer many questions facing health planners, policy analysts, and health care providers. For example, the file can be used in a microsimulation model in which the rules of various alternative medical assistance or national health insurance programs are sequentially applied to the information on each person to determine his or her eligibility, participation, and program utilization and expenditures. Aggregating these data over the sample individuals would give total program eligibility, participation, utilization, and costs for comparative analyses. Many equity and target efficiency questions can be addressed in this fashion.

The file can also be used in a broader context to estimate total national health care utilization and expenditures among the civilian, noninstitutionalized population and to give a breakdown of these estimates by state. The file could be aged by adjusting the population weights to equal the projected population in some future year and then used to project future health care utilization and expenditures. If the aging adjustments are made by state, age, and sex or by some other geographic and demographic breakdown, the effects of population shifts on health care utilization can be estimated.

For example, such an analysis could show the extent to which

health care demand might grow as the baby boom-era cohorts swell the ranks of the older population groups. An aged SIE/HIS file could also be used to estimate the growth in the share of health expenditures paid by the government as more and more people become eligible for Medicare benefits.

Thus, the success of the match shows that statistical matching of health care data to a large census file on an individual-record level is a viable method of creating a flexible database for health services research. The matching procedures are expensive and time-consuming. But when judged against the alternative of conducting a full-scale household survey, statistical matching may prove to be more efficient and more economical.

Furthermore, the procedure need not be restricted to building population-based files. It is easy to imagine a scenario where a file with records for different health facilities or health care services could be created from a statistical match.

ACKNOWLEDGMENTS

The author wishes to thank Dr. Harold Beebout and Dr. Constance Citro of Mathematica Policy Research, Inc., and Dr. Rodney Erickson and Dr. Edward Budd of The Pennsylvania State University, for comments on earlier drafts.

NOTES

1. The statistical match between the SIE and the 1976 HIS was performed by Mathematica Policy Research, Inc. and Social and Scientific Systems, Inc., under contract with the Health Care Financing Administration and the Office of the Assistant Secretary for Planning and Evaluation of the U.S. Department of Health and Human Services. For further information on the procedures used to create the merged database or on its availability, contact Mathematica Policy Research, Inc., 600 Maryland Avenue S.W., Suite 550, Washington, DC 20024.
2. The HIS actually consists of five record types—household records for each household interviewed, person records for each household member, and condition, hospital episode, and doctor visit records for each incidence of these among the household members. The SIE contains three record types—household records for each household interviewed, family records for each family in the household, and person records for each family member. Extracts of these files with only one record type per person were made for the match.
3. In record splitting, when the sample weights of matching records from the

two files are not equal, the record with the larger weight is split into two records, identical except for their sample weights. One of the new records is given the weight of the matching record in the other file and is matched with it. The other is given a weight equal to the difference in the original record weight and the matching record from the other file, and is replaced in the file of unmatched records for subsequent matching.

4. The actual output of the match is a series of files. One file contains the identification numbers of the SIE records and the matching HIS records. Another contains the original HIS records sorted by the identification number of the SIE records to which they were matched. From these files variables can be extracted and appended to the original SIE file.

REFERENCES

1. Alter, H. Creation of a synthetic data set by linking records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey. *Annals of Economic and Social Measurement* 373-94, April 1974.
2. Barr, R. S., and J. S. Turner. A New Linear Programming Approach to Microdata File Merging. *1978 Compendium of Tax Research.* Office of Tax Analysis, U.S. Department of Treasury. Washington, DC: U.S. Government Printing Office, 1978, pp. 131-49; Reply, pp. 152-55.
3. Boulding, W. Creation of a New Data Base Using the Survey of Income and Education and the Annual Housing Survey. Draft final report to Division of Housing and Demographic Analysis, U.S. Department of Housing and Urban Development. Washington, DC: Mathematica Policy Research, June 29, 1979.
4. King, J. A. The Distributional Impact of Energy Policies: Development and Application of the Phase I Comprehensive Human Resources Data System. Final report to Federal Energy Administration. Washington, DC: Mathematica Policy Research, June 1977.
5. Okner, B. Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement* 325-42, July 1972.
6. Radner, D. B. The Development of Statistical Matching in Economics. Paper presented at the 1978 meetings of the American Statistical Association, Social Statistics Section. San Diego, CA: August 16, 1978.
7. Radner, D. B. An example of the use of statistical matching in the estimation and analysis of the size distribution of income. *Review of Income and Wealth* 211-42, September 1981.
8. Radner, D. B. The Statistical Matching of Microdata Sets: The BEA 1964 CPS-TM Match. Unpublished Ph.D. dissertation, Yale University, May 1974.
9. Radner, D. B., and H. J. Muller. Alternative Types of Record Matching: Costs and Benefits. *Proceedings of the 1977 Meetings of the American Statistical Association, Social Statistics Section*, 1977, pp. 756-61.
10. Ruggles, N., and R. Ruggles. A strategy for merging and matching microdata sets. *Annals of Economic and Social Measurement* 353-72, 1974.
11. Ruggles, N., R. Ruggles, and E. Wolff. Merging microdata: Rationale,

practice, and testing. *Annals of Economic and Social Measurement* 407-28, 1977.
12. Springs, R., and H. Beebout. The 1973 Merged SPACE/AFDC File: A Statistical Match of Data from the 1970 Decennial Census and 1973 AFDC Survey. Final report to Social and Rehabilitation Service, U.S. Department of Health, Education, and Welfare. Washington, DC: Mathematica Policy Research, March 1976.
13. Fuchs, V. R. *Who Shall Live? Health, Economics, and Social Choice.* New York: Basic Books, 1974.
14. Feldstein, M. S. Econometric studies of health economics. In M. D. Intriligator and D. A. Kendrick (eds.). *Frontiers of Quantitative Economics*, Vol. II. Amsterdam: North Holland Publishing Co., 1974, pp. 377-434.
15. Kadane, J. B. Some Statistical Problems in Merging Data Files. *1978 Compendium of Tax Research.* Office of Tax Analysis, U.S. Department of Treasury. Washington, DC: U.S. Government Printing Office, 1978, pp. 159-71; Reply, pp. 177-79.
16. Hollenbeck, K. M., and P. Doyle. Distributional Characteristics of a Merged Microdata File. Paper presented at the annual meetings of the American Statistical Association. Washington, DC, August 13-16, 1979.
17. National Center for Health Statistics. *A Summary of Studies of Interviewing Methodology.* Vital and Health Statistics, Series 2, No. 69. U.S. Department of Health, Education, and Welfare, Public Health Service. Washington, DC: U.S. Government Printing Office, 1979, pp. 4-17.
18. Pappas, N. G. The Statistical Match Between the 1976 Health Interview Survey and the Survey of Income and Education. Final report to Health Care Financing Administration. Washington, DC: Mathematica Policy Research, November 1979.